DSSTox Field Definition File:

Carcinogenic Potency Database Summary Tables (CPDBRM, CPDBHA, CPDBDG, CPDBPR)

(last updated 15 October 03)

Description: Information in this file is intended to provide a minimum level of annotation to the DSSTox SDF files created for the Carcinogenic Potency Database Summary Tables (CPDBRM= rats and mice, CPDBHA= hamsters, CPDBDG= dog, CPDBPR= non-human primates) obtained from the CPDB Source website: http://potency.berkeley.edu/. For further explanation of Source-specific fields, a user is encouraged to consult the CPDB website, listed references, and documentation. Some modifications in fields (and allowable contents) were made to the original CPDB Summary Tables to facilitate use of the DSSTox SDF files in relational searching applications. All modifications are fully documented in the ***Comments*** section of the table below.

The first section of the table below lists and defines the ***DSSTox Standard Chemical Fields*** used in the CPDB SDF files. Any modifications in these fields, deviating either from the original Source data tables [1-3] or the ***Central List of DSSTox Standard Chemical Fields*** are noted in the ***Comments*** section. Following that section, all Source-specific fields in the DSSTox SDF files (i.e., CPDBRM, CPDBHA, CPDBDG, CPDBPR) are listed and defined. The ***DSSTox SDF*** column lists files in which the corresponding ***Field Name*** is present. All ***Units*** and ***Descriptions*** are extracted from Source reference materials unless otherwise noted. In some cases, modifications in ***Field Name*** and ***Allowable Values*** from the original data tables were made to facilitate creation and use of the DSSTox SDF files. All differences are noted in the ***Comments*** section. ***Allowable Values*** list allowable field entries occurring in CPDB SDF files separated by slashes for exclusive entries (i.e. cannot occur with another entry) and commas or spaces for non-exclusive entries (i.e. can occur with other values). These codes are defined and explained in the ***Description*** section; italicized note refers to the type of entry (e.g., *Text*). The pound symbol (#) indicates that the ***Allowable Values*** entry is a number. A pound symbol followed by a list of character options (e.g., # i, m, n, v) indicates that one or more footnote characters may follow the number entry; these are meant to provide additional information and are defined in the ***Description*** section. To minimize problems with import and export of SDF files, we avoid the use of punctuation and symbols in ***Allowable Values*** wherever possible; multiple entries in a single field (e.g., adr cli eso) are separated by a single space in the SDF. Upper and lower cases in ***Allowable Values*** text entries are used only for emphasis, and not alone to distinguish separate meaning, except in the case of SMILES codes, which are case-sensitive.

For the CPDBRM and CPDBHA summary tables, information has been added to the DSSTox SDF files to alert a user of possibly redundant chemical information in the chemical structure fields, including replicate CAS number entries in the database (these corresponding to different technical grades or formulations in most cases), and to cases where 2D structures or simplified parent forms of salts or complexes have replicate entries in the database (e.g., cis and trans isomers of the same chemical, or Na and K salts of the same chemical). Two fields, **ChemNote** and **ChemCount**, contain information pertaining to replicate CAS, parent structure, or 2D structure entries and can be used to find and count the number of unique chemical substances, parent structures, or 2D structures in the database. Note that the use of this term “replicate” in the Standard Chemical Fields pertains only to chemical information in the database that may require extra consideration before use in structure-activity modeling; replicate records in the CPDB always correspond to different experimental studies and biological results.

Source Website: The CPDB, from which the Summary Tables are derived, is available in several formats at http://potency.berkeley.edu/

Source Contact: Please contact Lois Swirsky Gold for questions pertaining to the content of the CPDB Summary Tables; email: cpdb@potency.berkeley.edu
Please contact DSSTox Support for questions or comments pertaining to the DSSTox CPDB SDF files.

Main Citations: Publications reporting use of DSSTox SDF files for the CPDB Summary Tables are asked to list the full DSSTox file name(s), including date stamp, and to cite as primary reference the following 4 citations:

http://potency.berkeley.edu/

Gold, L.S., Slone, T.H., Ames, B.N., Manley, N.B., Garfinkel, G.B., and Rohrbach, L. (1997) Carcinogenic Potency Database. In: Gold, L.S., and Zeiger, E., Eds. Handbook of Carcinogenic Potency and Genotoxicity Databases. Boca Raton, FL: CRC Press, pp. 1-605. http://potency.berkeley.edu/CRCbook.html

Gold, L.S., Manley, N.B., Slone, T.H., and Rohrbach, L. (1999) Supplement to the Carcinogenic Potency Database (CPDB): Results of animal bioassays published in the general literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996. *Environ. Health Perspect.* 107 (Suppl. 4): 527-600. http://ehpnet1.niehs.nih.gov/docs/1999/suppl-4/toc.html

Gold, L.S., Slone, T.H., Ames, B.N., Manley, N.B.,Summary Table of Chemicals in the Carcinogenic Potency Database: Results for Positivity, Potency (TD_{50}), and Target Sites: http://potency.berkeley.edu/chemicalsummary.html

SDF Development Notes:
Each DSSTox SDF file contains a single **Structure** field whose entry corresponds to the **StructureShown**, **CAS**, **SMILES**, **Formula**, and **MolWeight** fields. The main DSSTox SDF files represent the actual tested form of the chemical in the **Structure** field (see ***Description*** below), including complexed molecular entities and salt counter ions in all cases. An additional DSSTox "Defined Organic Parent" SDF file (CPDBRM_DOP) is offered for download only for the largest CPDBRM file, for specialized use in Structure-Activity Relationship (SAR) modeling applications. This DOP file contains no inorganics, organometallics, or mixtures, and all defined organic salts and complexes are stripped of counter-ions and complexed molecular entities and converted to a simplified parent representation in the **Structure** field. The **StructureShown** entry for these compounds is "simplified to parent", with corresponding changes in the **CAS**, **SMILES**, **Formula**, and **MolWeight** field entries. These "simplified to parent" structures are represented in neutralized (protonated) form wherever possible [exceptions include quaternary ammonium and pyridinium ions, which are represented as positively charged (N+) stripped of counter ions, and nitro compounds, which are represented in the charge-separated form, i.e. N+(=O)(O-)]. In the DOP file, both a **CAS_TestedForm** and **SMILES_TestedForm** field are included to allow a user to refer back to the original CAS and SMILES of the tested form of the chemical (i.e., salt or complex). The remaining field contents of the DOP file are identical to that of the main CPDBRM SDF for the subset of "defined organics". Users should be aware that most commercial chemical relational database applications automatically insert one or more structure identifier fields upon import or export of an SDF file (e.g., FW or Mol_ID), fields that may augment or duplicate one or more of the DSSTox Standard Chemical Fields. Also, since the proper ordering of fields upon SDF import into many applications requires a non-blank entry in each field of the first database record, the word "blank" is entered in each empty text field in Record 1 for this purpose; this word can be deleted from Record 1 fields after SDF import is complete.

As an MS Word document, the following table is best viewed onscreen using either Normal or Web Layout View in Landscape page orientation. Page breaks have been inserted in both the MS Word and PDF versions of this table for optimal page layout view and printing; note that this results in the **Target Sites** *row being split over 2 pages.*

Field Name	***DSSTox SDF***	***Units***	***Allowable Values***	***Description***	***Comments***
DSSTox Standard Chemical Fields					
Structure	All		*Molecule*	Two-dimensional graphical representation of molecular structure. Form of structure is identified in the **StructureShown** field and always corresponds to the fields: **CAS**, **SMILES**, **MolWeight**, **Formula**. Structure entry is *blank* when **SubstanceType** entry is "mixture or unknown".	
Structure Shown *(no spaces)*	All		tested form/ simplified to parent/	Identifies form of graphical 2D structure displayed in the **Structure** field. Field entry is "tested form" for all records in main SDF. For DOP file, entry is "simplified to parent" for "defined organic" salts and complexes.	
Formula	All		*Text*	Empirical formula of displayed **Structure**.	
MolWeight	All	amu	#	Molecular weight of displayed **Structure**.	

CAS	All		NOCAS/ ######-##-#/	Chemical Abstracts Service (CAS) Registry number of displayed **Structure**, formatted 000000-00-0, corresponds to **StructureShown** and **SMILES**. "NOCAS" indicates CAS number was not found.	
SMILES	All		*Text*	SMILES molecular text code of displayed **Structure**, corresponding to **StructureShown** and **CAS**.	
DSSTox_ID	All		#	Sequential ID number assigned to each record in database, values range from 1 to n= total #records. When accompanied by full DSSTox filename, provides unique record identifier.	
DSSTox_ FileName	All		*Text*	Full DSSTox SDF standard file name without .sdf extension. Field entry will be updated whenever new version or revision of SDF database is generated.	e.g., CPDBRM_v1a_1354_15Oct03 --or-- CPDBRM_DOP_v1a_1189_15Oct03
ChemName	All		*Text*	Common or trade name of chemical listed in original Source data table, corresponds to the original tested form of the chemical or substance, if known.	Table of chemical name synonyms available from CPDB Source website. Notation "s" following chemical name is used in original CPDB Rats and Mice Summary Table to indicate results are available for other species in the CPDB Summary Tables; a new field entitled "Other Species" is included in CPDBRM SDF to eliminate this footnote.
Substance Type *(no spaces)*	All		defined organic/ inorganic/ organometallic/ mixture or unknown/	Nature of chemical or substance: "defined organic" = defined chemical structure containing carbon but not organometallic, i.e. with no metal or metalloid other than simple salt alkali (I) or alkali earth (II) metals; "inorganic" = defined chemical structure containing no carbon; "organometallic" = operationally defined as a chemical structure containing carbon and any metal or metalloid other than alkali (I) or alkaline earth (II) metals (such as Na, K, Mg, Ca) that occur in simple salts; organometallics always labeled "complex" in **TestedForm** field; "mixture or unknown" could represent mixture, ill-defined, or unknown substance type, and has no entry in **Structure**, **StructureShown**, **TestedForm**, and **SMILES** fields.	Information inferred from "Chemical" and "CAS" listed in original CPDB Summary Tables.
TestedForm	All files containing tested substances		parent/ salt, complex/	Tested form of chemical inferred from **ChemName**, **CAS** and **Structure**, with DSSTox operational definitions as follows: "parent" = single chemical entity, without counter ions or complexed chemical entities; "salt" = simple ionic salts of alkali (I) or alkaline earth (II) metals (such as Na, K, Mg, Ca) or halides (e.g., Cl, Br); "complex" = any compound with associated acid, base, or hydrate, or any organometallic.	Original tested form of chemical inferred from "Chemical" and "CAS" columns listed in original CPDB Summary Tables.

AddToParent	All files containing salts or complexes		*Text*	For **SubstanceType=**"defined organic" and **TestedForm=**"salt" or "complex", entry specifies salt counter-ions or complexed entities (e.g., Na, K, HCl, Cl, H2O, Ca , H2SO4, acetate, etc.) that are removed when **StructureShown**="simplified to parent" in DOP file; "bis" signifies parent structure occurs twice in complex.	
ChemNote	All		*Text,* quaternary ammonium, stereochem, replicate CAS, replicate 2D, replicate parent, *etc.*	Note related to nature of chemical in exceptional cases when, e.g., uncertainty exists in **ChemName** or **CAS**, parent structure is "quaternary ammonium" ion, mixture characteristics are known, "stereochem" information is known (e.g., *cis, trans, E,Z*, R, S), etc. Note also indicates instances of "replicate" information pertaining to CAS or structure: replicate CAS: same CAS number replicate 2D structure: geometric or stereoisomers replicate parent structures: salt or complex of same parent structure See also **ChemCount** field for counter of all replicate sets.	Use of the term "replicate" does not necessarily imply identical chemicals or substances, but is intended to alert a user to possibly redundant chemical information that may not be distinguished in chemical property calculation modules or structure-activity models.
ChemCount	Files containing replicate CAS and structure information (CPDBRM)		1/ # of #*total*	Counter field specifying instances of replicates in the database. Entry is "1" by default. If replicates exist, entry is a counter number (1,2,3, etc) followed by "of" and the total number of replicates in the replicate set., e.g., 1 of 3, 2 of 3, 3 of 3 for a set consisting of 3 replicates.	**ChemNote** field entry will indicate type of replicate, whether CAS, 2D structure, or parent structure. Search for "1" to find all unique entries, i.e. excluding all types of replicates beyond the first instance. Search for "*of" (* is application-dependent wildcard symbol) to find all instances of replicate information. Search for "1 of" to find number of replicate sets. **ChemCount** field occurs in CPDBRM but not in the other CPDB SDF files.
CAS_TestedForm	Files containing simplified structures, e.g., DOP		NOCAS/ ######-##-#/	CAS of actual tested form of chemical, formatted **000000-00-0**, differs from **CAS** field entry only when a simplified form of chemical is represented in the **Structure** field. "NOCAS" indicates CAS number was not found.	Field occurs only in DOP file and entry differs from **CAS** only for salts and complexes when **StructureShown**="simplified to parent".
SMILES_ TestedForm	Files containing simplified structures, e.g., DOP		*Text*	SMILES molecular text code of actual tested form of chemical, differs from **SMILES** field entry only when a simplified form of chemical is represented in the **Structure** field. Corresponds to **CAS_TestedForm**.	Field occurs only in DOP file and entry differs from **SMILES** only for salts and complexes when **StructureShown**="simplified to parent".

CPDB Source-Specific Fields					
SAL CPDB	CPDBRM CPDBHA CPDBDG CPDBPR	*None*	pos/ neg/ NE/	A chemical is classified within the CPDB as mutagenic, i.e. "pos", in the *Salmonella* assay if it was evaluated overall as either "mutagenic" or "weakly mutagenic" by Zeiger [4] or as overall "positive" by the EPA Gene-Tox Program [5,6]. All other chemicals evaluated for mutagenicity by these two sources are reported as "neg". "NE" for a chemical indicates no evaluation of mutagenicity from either source. This is a summary mutagenicity determination in the CPDB Summary Table that is based on overall evaluations (not strain-specific for *Salmonella*) from two sources of overall evaluations, using the above rule.	This field is titled "Salmonella" in the original CPDB Summary Tables; symbol entries appearing in this field were converted to the following DSSTox text equivalents: "+" = pos (positive) "-" = neg (negative) "." = NE (no evaluation)
TD50 Rat/ **TD50 Mouse** **TD50 Hamster** **TD50 Dog** **TD50 Rhesus/** **TD50 Cynomolgus/** **TD50 Tree Shrew/** **TD50 Bush Baby**	CPDBRM CPDBHA CPDBDG CPDBPR	mg/kg/ day	NP/ ND/ IA/ # i, m, n, v/	TD50 is a standardized quantitative measure of carcinogenic potency (analogous to an LD50) and is computed in the CPDB for each species/sex/tissue/tumor type for each experiment. TD50 is defined as: "that dose-rate in mg/kg body wt/day which, if administered chronically for the standard lifespan of the species, will halve the probability of remaining tumorless throughout that period" [4-6]. In the original CPDB Summary Tables, a TD50 (#) is reported for a chemical in each species with a positive evaluation of carcinogenicity in at least one test. If there is only one positive test on the chemical in the species, then the most potent TD50 from that test is reported. If more than one experiment is positive, the reported TD50 is the harmonic mean of the most potent TD50 values from each positive experiment in the species (denoted with an "m" as defined below) [2,3,7-9]. If a numerical value of the TD50 is listed, it may be accompanied by one or more of the following footnote indicators: i = Intraperitoneal or intravenous injection are the only routes of administration with positive tests in the CPDB. m = More than one positive test in the species in the CPDB; in this case the reported TD50 is the harmonic mean of those positive results [7]. n = No results evaluated as positive for this species in the CPDB are statistically significant ($p<0.1$). v = Variation is greater than ten-fold among statistically significant ($p<0.1$) TD50 values from different positive experiments in the species. If a TD50 value is not reported for the listed chemical and species, the field entry will be one of the following: "NP" indicates no positive results, "ND" indicates no data available for that chemical and species or species-sex category, "IA" indicates NCI/NTP bioassays were the only available experiments and both sexes in the species were evaluated as inadequate.	These fields appear under the column headings "Harmonic mean of TD50 (mg/kg/day)" in the original CPDB Summary Tables. The field names have been shortened and units eliminated to simplify and lessen the probability of error in the DSSTox SDF files. Abbreviations from the original CPDB Summary Tables appearing in the TD50 fields were converted to the following DSSTox text equivalents: "-" = NP (no positive results) "." = ND (no data) "I" = IA (inadequate NCI/NTP bioassays) In the CPDBPR, TD50 results were reported for a single chemical in the case of both the Tree Shrews and Bush Babies. For these species, only the single result is listed and all other chemical fields are left *blank*, indicating no data are in the CPDB.

Target Sites Rat Male/ **Target Sites Rat Female/** **Target Sites Rat Both/** **Target Sites Mouse Male/** **Target Sites Mouse Female/** **Target Sites Mouse Both/**	CPDBRM	*None*	NP/ ND/ IA/ adr bon cli eso ezy gal hag hmo kid lgi liv lun meo mgl mix mvc nas nrv orc ova pan per	Tumor target sites are reported using mnemonic codes as follows: adr = adrenal gland; bon = bone; cli = clitoral gland; eso = esophagus; ezy = ear/Zymbal's gland; gal = gall bladder; hag = harderian gland; hmo = hematopoietic system; kid = kidney; lgi = large intestine; liv = liver; lun = lung; meo = mesovarium; mgl = mammary gland; mix = mixture; myc = myocardium; nas = nasal cavity (includes tissues of the nose, nasal turbinates, paranasal sinuses and trachea); nrv = nervous system; orc = oral cavity (includes tissues of the mouth, oropharynx, pharynx, and larynx); ova = ovary;	These fields appear under the column headings "Rat target sites", "Mouse target sites" and "Target sites" in the original CPDB Summary Tables. The field names have been separated and standardized in the DSSTox SDF files. Abbreviations from the original CPDB Summary Tables appearing in the Target site fields were converted to the following DSSTox text equivalents: "-" = NP (no positive results) "." = ND (no data) "I" = IA (inadequate NCI/NTP bioassays) In the CPDBPR, Target Site results were reported for a single chemical in the case of both the Tree Shrew and Bush Baby. For these species, only the single result is listed and all other chemical fields are left *blank*, indicating no data are in the CPDB. In the CPDB Summary Tables for Rats, Mice, and Hamsters, the notation (B) is used in conjunction with a species-sex target site, e.g., "adr (B)" to indicate that a study reported experimental results only for the two sexes combined. This notation would appear for the target site under both Male and Female headings unless one of two conditions held: (1) positive results were available for the target site in the other sex of the species; or (2) negative results were reported for the other sex of the species. A "B-" would be listed alone (i.e.,
Target Sites Hamster Male/ **Target Sites Hamster Female/** **Target Sites Hamster Both**	CPDBHA				
Target Sites Dog	CPDBDG				
Target Sites Rhesus/ **Target Sites Cynomolgus/** **Target Sites Tree Shrew/** **Target Sites Bush Baby**	CPDBPR				

			pit pre pro ski smi spl sto sub tba tes thy ubl ute vag vsc H	pan = pancreas; per = peritoneal cavity; pit = pituitary gland; pre = preputial gland; pro = prostate; ski = skin; smi = small intestine; spl = spleen; sto = stomach; sub = subcutaneous tissue; tba = all tumor bearing animals; tes = testes; thy = thyroid gland; ubl = urinary bladder; ute = uterus; vag = vagina; vsc = vascular system. "H" = A mix of carcinomas of the ear duct, Zymbal's gland, oral cavity or nasal cavity were combined by single author, Maltoni, in his category "Head cancers", which he reports as induced by the chemical. Target sites are reported for each sex-species group with a positive result in the CPDB. Target sites are identified on the basis of a positive author's opinion for the particular site, in any experiment in the species-sex or species, using all results from both the general literature and NCI/NTP bioassays. If a chemical has two or more target sites listed, the results may be from different experiments, and a single site may be from more than one experiment, as well [2,3]. CPDB data organized by target site have been published by Gold et al. [3]. If a tumor target site code is not reported for the listed chemical and species, the field entry will be one of the following: "NP" indicates no positive results, "ND" indicates no data available for that chemical and species-sex category, equivalent to *blank* entry in CPDBPR. "IA" indicates NCI/NTP bioassays were the only available experiments and were evaluated as inadequate.	without target site codes) if negative results were only found in a study where the two sexes were combined. Since these results are not resolved to species-sex, we have created separate fields in the CPDBRM and CPDBHA SDF files (Rat Both Target Sites, Mouse Both Target Sites, Hamster Both Target Sites) to accommodate this information. Non-*blank* entries in these fields only occur under the same conditions reported in the CPDB Summary Tables, i.e. if combined-sex results are available and species-sex results are unavailable for either sex. The Target Sites in CPDBPR are reported for species with sexes combined. In the CPDB Summary Tables, four chemicals in rats include the footnote "h" to denote a positive opinion for a consolidated site "head tumors", by only a single research group. Maltoni et al. evaluated head as a target site for a combination of carcinomas at various sites in the head (ear duct, Zymbal's gland, nasal cavity, or oral cavity). When this result has confirming results from other studies for either nas or orc target sites in the head, it is reported as nash and orch in the CPDB Summary Table. When no other positive results are reported, the result is designated "+h" in the CPDB Summary Table [4]. We have replaced both "h" and "+h" with "H" in the CPDBRM file. This occurs either alone, or following a target site, as in " nas H" or "orc H".

Other Species	CPDBRM CPDBHA CPDBDG CPDBPR	*None*	CPDBRM CPDBHA CPDBDG CPDBPR	Non-*blank* entry indicates that data for this chemical exists in other CPDB Summary Tables and corresponding DSSTox SDF files, i.e., CPDBRM for rat and mouse CPDBHA for hamster CPDBDG for dog CPDBPR for rhesus, cynomolgus, tree shrew or bush baby	This field replaces use of "s" footnote following chemical name (under Chemical heading) in original CPDB Summary Tables. If more than one file name is listed, this indicates data exists for this chemical in multiple species and therefore in multiple CPDB Summary Tables and corresponding SDF files.

Additional CPDB references:

A list of citations and text files to 100 papers by the Carcinogenic Potency Project: http://potency.berkeley.edu/listofpubs.year.html

1. Gold, L.S. and Zeiger, E., Eds. (1997). Handbook of Carcinogenic Potency and Genotoxicity Databases. Boca Raton, FL: CRC Press. http://potency.berkeley.edu/CRCbook.html

2. Gold, L.S., T.H. Slone, and B. Ames (1999) Summary of Carcinogenic Potency Database by Chemical. http://potency.berkeley.edu/chemicalsummary.html

3. Gold, L.S., N.B. Manley, T.H. Slone, and J.M. Ward (2001) Compendium of chemical carcinogens by target organ: Results of chronic bioassays in rats, mice, hamsters, dogs and monkeys. *Toxicol. Pathol.* 29: 639-652. http://potency.berkeley.edu/text/ToxicolPathol.pdf

4. Zeiger, E. (1997) Genotoxicity Database. In: Gold, L.S., and Zeiger, E., Eds. Handbook of Carcinogenic Potency and Genotoxicity Databases. Boca Raton, FL: CRC Press, pp. 687-729. http://potency.berkeley.edu/CRCbook.html

5. Kier, L.E., D.J. Brusick., A.E. Auletta, E.S. Von Halle, M.M. Brown, V.F. Simmon, V. Dunkel, J. McCann, K. Mortelmans, M. Prival, T.K. Rao, and V. Ray (1986) The *Salmonella typhimurium*/mammalian microsomal assay: A report of the U.S. Environmental Protection Agency Gene-Tox Program. *Mutat. Res.* 168: 69-240.

6. Auletta, A.E., Personal communication (with L.S.Gold).

7. Peto, R., M.C. Pike, L. Bernstein, L.S. Gold, and B.N. Ames (1984) The TD_{50}: A proposed general convention for the numerical description of the carcinogenic potency of chemicals in chronic-exposure animal experiments. *Environ. Health Perspect.* 58: 1-8.

8. Sawyer, C., R. Peto, L. Bernstein, and M.C. Pike (1984) Calculation of carcinogenic potency from long-term animal carcinogenesis experiments. *Biometrics* 40: 27-40.

9. Gold, L.S., T.H. Slone, and L. Bernstein (1989) Summary of carcinogenic potency (TD_{50}) and positivity for 492 rodent carcinogens in the Carcinogenic Potency Database. *Environ. Health Perspect.* 79: 259-272.